Midterm Project Requirements

Submission Deadline: Monday, March 3rd, at 12:00 PM (Noon) via iCollege

Project Overview:
For the midterm project, each group may select one of two available datasets:

- FDD Item 5 Dataset (relatively easier)
- FDD Item 12 Dataset

Each group is required to submit both:

1. A written report (minimum of five pages)
2. A presentation deck (approximately 15 slides)
3. A python/Jupyter Notebook file

Each group should submit only one set of results. Individual submissions from group members are not required, as this would lead to duplicate submissions.

Grading Criteria

1. Coding Component (60 points)

Each group must complete the following tasks:

1. Select one dataset from the two provided options. (Note: The Item 5 dataset is slightly less complex.)
2. Choose one file from the selected dataset and generate three word cloud visualizations using: *(10 points)* ***(Use any one file among 50 files)***
    - Bag of Words (BoW)
    - N-Gram
    - Term Frequency-Inverse Document Frequency (TF-IDF)

3. Apply at least at least one frequency-based method and one embedding method, conduct topic modeling, and provide an interpretation of the results. *(10 points)* *(Use all 50 files)*
4. Question Answering Tasks: *(40 points)(Use all 50 files)*
    - For the Item 5 dataset: Extract the initial fee from each file.
    - For the Item 12 dataset: Analyze whether the franchisee is offered an exclusive territory.

5. The final grading for this section will be based on the accuracy of the question-answering results.

2. Report Requirements (25 points)

1. Clearly explain the task and provide a summary of the dataset, including a description of the variables, their dimensions, and data types. *(10 points)*
2. Present and interpret the word cloud visualizations and topic modeling results. *(10 points)*
3. Display the question-answering results in a structured table and report the accuracy. *(5 points)*

3. Presentation (10 points)

Each group will be randomly asked two questions during their presentation, with each question worth 5 points.

4. Formatting Requirements (5 points)

1. Use your business question as the title of both your report and presentation. *(1 point)*
2. Include your group number and all group members' names (ordered alphabetically by last name) on the first slide and the first page of your report. *(1 point)*
3. Minimize the use of direct Python code screenshots in your slides and report. Instead, present outputs in tables and visualizations. *(3 points)*

Please ensure that your submission meets all the outlined requirements. Let me know if you have any questions.

Update:

Hello everyone,

I have uploaded the answers for Item 5 and Item 20. Please use them as the reference standard to assess the accuracy of your 'Question Answering' results.

- **Item 5: If your answer matches the 'True Initial Fee', it is correct. If it matches any value in 'Alternative Answer', it is half correct. If it does not match either, it is incorrect.**
- **Item 12: If your extracted text matches 'Original Decision', it is correct; otherwise, it is incorrect.**

Each group needs to submit a CSV file that compares your Question Answering results with the provided answers. Additionally, when I rerun your submitted program, it should produce the same results as yours.

Each group will have about 15 minutes to present, and I will ask two questions during each presentation.

Update:

Hi Everyone,

After receiving an inquiry from one group, I would like to clarify the process of topic modeling and address a correction for one of the questions.

1. **Topic Modeling:**
   Each file should be treated as an individual document. Your task is to perform clustering (topic modeling) on a set of 50 files and generate results.
   For example, if you choose to divide the files into three clusters, each file must be assigned to one of the three topics (e.g., Topic 1, Topic 2, or Topic 3). The output should specify which files belong to each cluster. For instance:
   - **Cluster 1: File 1, File 10, File 12, etc.**
   - **Cluster 2: File 2, File 5, File 15, etc.**
   - **Cluster 3: File 3, File 7, File 20, etc.**
2. **Ensure that every file is assigned to exactly one cluster.**
3. **Question Answering (Item 5): For the file *"8 Domino's"*, both "$10,000" and "$0 - $10,000" will now be considered correct answers. I appreciate those who pointed this out.**